

Sesgo, error y aleatoriedad en la investigación

(Bias, error and randomness in investigation)

Badii, M.H., A. Guillen, M. García-Martínez & J.L. Abreu
UANL, San Nicolás, N.L., México

Abstract. In this paper types of bias, error (systematic as well as random) and randomness are noted. Types of statistical models and their components along with their application for prediction are described. Furthermore, data quality and a metric (COETE) to ensure collection of correct data are mentioned.

Keywords: Bias, COETE, data quality, error, randomness.

Resumen. En este trabajo se mencionan los tipos de sesgo, errores tanto sistemático como aleatorio y el componente al azar. De forma breve se notan las clases de los modelos estadísticos, sus componentes y sus aplicaciones para la noción de la predicción. Además se describen la calidad de datos, y una métrica (COETE) para seleccionar datos correctos que garanticen su calidad.

Palabras clave: Al azar, calidad de datos, COETE, error, sesgo

Introducción

¿Por qué pensamos que la muerte llega más rápido a otros que a nosotros? ¿Por qué no ponemos atención a los valores reales de la probabilidad cuando al temor de viajar con avión versus el automóvil? (Nociones básicas de la ciencia de la probabilidad).

La ciencia de la estadística constituye la base fundamental para la toma de decisiones. En este espacio se tratan los siguientes temas esenciales (Kepner y Robinson; Badii et al, 2006; 2007; Badii et al, 2009; Wheelan; 2014; Stigler, 2017).

- 1.- Definición
- 2.- Tipos de los errores
- 3.- Modelos
- 4.- Calidad de los datos

Definición

Definición Conceptual:

La estadística se trata de medir la probabilidad de ocurrencia de las cosas (objetos, eventos, procesos, fenómenos) y conceptos en el espacio y el tiempo y validar éstos niveles ya medidas (Badii et al, 2009).

Definición Operacional:

Descriptivo: lo cual se trata de tres actividades:

- A.- Coleccionar los datos o el muestreo. Este es la primera parte operacional de la estadística y por tanto, la calidad de los datos define la calidad de los resultados y las conclusiones derivadas de ellas.
- B.- Organizar los datos coleccionados, por ejemplo en el orden de magnitud.
- C.- Presentar los datos ya organizado, en forma de tablas, figuras, etc.

Definición Inferencial: lo cual conduce a la toma de decisión, y consiste en tres actividades:

- A.- Analizar los datos: convertir los datos en las estimaciones.
- B.- Interpretar los datos: aplicar los resultados del análisis a la vida, es decir, para que sirven los resultados obtenidas.
- C.- Publicar los datos: hacer que los resultados lleguen al público por diferentes medios, pero de preferencia por medio de las publicaciones serías científicas. Es relevante recalcar que un trabajo no publicado es igual a un trabajo no realizado.

La noción y las clases de errores

Los errores se producen por dos factores; la aleatoriedad y las causas sistemáticas (Badii et al, 2009; Kraus, 2012).

El fenómeno de la aleatoriedad.

El error aleatorio se puede disminuir por los siguientes factores:

- A.- Aleatorización de los datos. Al aleatorizar los datos, en realidad evitan el sesgo.
- B.- Número de repeticiones, este nos conduce al tamaño óptimo de la muestra, hay que señalar que al tener más de una repetición. Esto se permite cuantificar el error del experimento.

C.- Control de las condiciones o las variables que no están de interés y que pueden causar ruido o error. Esto nos ayuda a mitigar la magnitud del error. Existen diseños experimentales, como por ejemplo, el Diseños de Bloques Al azar, el Diseño de Cuadro, Latino, El diseño de Parcelas Divididas, el Diseño Factorial y el Diseño de ANCOVA (Análisis de Covarianza), los cuales son herramientas estadísticas a nuestra disposición que permiten controlar las variables no deseadas.

Causas sistemáticas.

Estos errores sistemáticos se producen por o ser humano o las maquinas, materiales o equipos utilizados en la investigación. El ser humano comete este tipo de error debido a descuido, o cansancio. Los errores sistemáticos generados por las maquinas, equipos o materiales pueden generarse debido al tiempo del uso, vibraciones, mala calidad del mantenimiento, etc. Una manera de reducir este tipo del error es por medio del Sistema de Aseguramiento de Calidad (incluyendo la ingeniería de calidad y el control de calidad), a través de auditorías tanto en procesos y en forma retrospectiva todo esto basado en capacitación y la mejora continua (Kepner y Robinson, 1988; Keppel, 1991; Kirk, 1982; Outwaite y Rutherford, 1955; Morris, 1999).

Ahora bien, en cualquiera situación donde las personas tienen que tomar decisiones, invariablemente, uno puede cometer uno de los dos tipos de errores (Badii et al, 2009). **A.-** Error tipo I o alfa o el falso positivo, y **B.-** Error tipo II o beta o el falso negativo.

Para demostrar estos dos tipos de errores se puede observar la Tabla 3.1.

Tabla 3.1. Clases de errores asociados a la toma de decisión.		
Decisión	Ho: El sujeto es inocente.	Ha: El sujeto no es inocente
	Hipótesis nula (Ho)	
	Cierto	Falso
Rechazar	Error tipo I = α = Falso positivo	No error (Potencia estadística, $1 - \beta$)
Aceptar	No error (Nivel de confianza, $1 - \alpha$)	Error tipo II = β = Falso negativo

Según los arreglos en Tabla arriba, suponemos que este en juego la inocencia o la culpabilidad de una persona sospechosa ante un juez. En este contexto la hipótesis nula (H_0) declara que “el sujeto es inocente”, mientras que la hipótesis alterna (H_a) dice que el sujeto no es inocente sino culpable. Ahora bien, el Juez después de analizar todos los documentos y las evidencias en favor y en contra de la culpabilidad del sujeto tiene que llegar con una decisión de dictaminar la inocencia o la culpabilidad del sujeto.

Caso de error tipo I: Suponemos que la hipótesis nula es correcta y el sujeto es realmente inocente, Sin embargo, siendo el juez finalmente un ser humano, puede cometer uno de los dos errores arriba mencionados. El juez puede rechazar la noción de la inocencia y por tanto, declara de forma errónea culpable a una persona inocente. En ese caso el juez está cometiendo el error de tipo I o alfa o falso positivo. En las ciencias sociales el nivel usual del error tipo I es el 5%. Por tanto, $1 - 0.05$ es igual a 0.95 o 95% y este es el nivel de confianza con el cual el juez hace el dictamen (y con todo error involucrado).

Caso de error tipo II: Suponemos que la hipótesis nula es incorrecta y el sujeto es realmente culpable, Sin embargo, siendo el juez finalmente un ser humano, puede cometer uno de los dos errores arriba mencionados. El juez puede aceptar la noción de la inocencia y por tanto, declara de forma errónea inocente a una persona culpable. En ese caso el juez está cometiendo el error de tipo I o beta o falso negativo. En las ciencias sociales el nivel usual del error tipo II es el 10%. Por tanto, $1 - 0.10$ es igual a 0.90 o 90% y este es el nivel de potencia estadística con el cual el juez asegura el dictamen (y con todo error involucrado). Este ejemplo ilustra que al tomar decisiones, siempre existe la probabilidad de cometer uno de los dos tipos de errores mencionados, y el costo de cometer el error en estos casos está directamente correlacionado con si el asunto bajo el interés es un asunto de salud, o emocional y financiero (Badii et al, 2009).

Modelos

Un modelo es un conjunto de supuestos sobre lo que se estudia. Estadísticamente, un modelo es una ecuación matemática para representar la relación entre variables. Esta relación se base en pocos números llamados *parámetros*.

Tipos de Modelo

Existen muchos diferentes modelos, sin embargo, los más simples y flexibles son los modelos lineales (Pezzullo, 2013; Celis y Labrada, 2014; Soka y Rohlf, 1969; Samuels et al, 1991).

Modelo Lineal: En este tipo de modelo, los Parámetro se interpretan como *Constantes Fijos*. En esencia el modelo lineal trata de describir e explicar una variable respuesta (dependiente) a partir de las variables predictores (independientes). Hay que notar que en ciencias sociales, la salud, etc, las variables respuestas dependen normalmente de un conjunto de variables predictores. En término general, la *respuesta* depende de los factores *Tomados en cuenta* y de los factores *no tomados en cuenta* (Kepner y Robinson, 1988)

Utilidad de Modelos Lineales: Los modelos lineales sirven para a) Medir variables continuas, categóricas, interacciones y b) Estimar las coeficientes (los β 's), este requiere conocer la distribución de "Y". *Por tanto, un Modelo Lineal Tiene 2 partes:* a) una parte visible lo cual es la ecuación, y b) La parte invisible, es decir la distribución de "Y", y las restricciones de éste distribución sobre parámetros de la ecuación.

Variable observada del factor dependiente = Efecto de factores tomados en cuenta + Efecto de factores no tomados en cuenta

Los de lado derecha explican la parte izquierda, pero no hay control sobre el segundo factor, que escapa al investigador, Por tanto:

Variable observada del factor dependiente = Efecto de factores constantes + Efecto de factores tomados en cuenta + Efecto de factores no controlados o error

$$Y_i = \beta_0 \text{ (común a todos sujetos)} + \beta_i X_i \text{ (efecto de factores Independiente)} + E_i$$

Debido a que: $\mu_i = \beta_0 + \beta_i X_i$,

Por tanto: $Y_i = \mu_i + E_i$

Variable que desea explicar = La parte que explica + error

Ejemplo: Efecto IQ sobre rendimiento académico:

Rendimiento esperado = Rendimiento promedio Común a todos los sujetos + Rendimiento Debido al IQ de cada sujeto.

Rendimiento observado (Y_i) = Rendimiento esperado (μ_i) + error (E_i)

Donde, $\mu_i = \beta_0 + \beta_i X_i$, es decir, Rendimiento Esperado = Rendimiento Promedio Común a todos los sujetos + Rendimiento debido al IQ de cada sujeto.

Para variables Continuas: β mide cambio en “Y” asociada al cambio unitario en “X”.

Para variables Categóricas: β mide cambio en “Y” asociada al cambio de categoría (nivel) en “X”. En resumen:

Variables continuas: $Y_i = \beta_0 + \beta_i X_i + E_i$ & $\mu_i = \beta_0 + \beta_i X_i$, por tanto: $Y_i = \mu_i + E_i$

Variables categóricas: $Y_{ij} = \mu + \alpha_i + E_{ij}$ & $\mu_{ij} = \mu + \alpha_i$, Por tanto: $Y_{ij} = \mu_{ij} + E_{ij}$

Componentes de un Modelo Lineal

Si la “Y” es Cuantitativa: La “Y” tiene una distribución “Normal” y con varianzas constantes para cada valor de “X”.

Si Y es Categórica (presencia-ausencia, acierto-error, vivo -muerto): Cada observación es un ensayo de Bernoulli y el número de aciertos en “n” ensayos tiene una distribución Binomial.

Si Y es 1 recuento (número de choque en un año, número de recaídas en un mes): Se usa una distribución que maneja los números enteros no-negativos como Poisson.

Componente sistemático: las “X’s”

A.- Variables de “X” son con efectos fijos.

- B.-** Cada variable “X” contribuye al pronóstico con una tasa lineal igual a β_i .
- C.-** La variable “X” es: cuantitativo, categórico, transformado (un modelo polinomial cuadrático; $X^2 = X_1^2$; y para un modelo con el efecto de interacción: $X_3 = X_1 * X_2$)

Componente “Función de enlace”

- A.-** Distribución Normal: Su parámetro natural es su propia media, por tanto, se usa una función de enlace “Identidad”: $g(\mu_i) = \mu_i = \beta_0 + \sum \beta_i X_i$. Se usa este tipo de distribución para ANOVA, Regresión (pronósticos en la misma métrica que la “Y”).
- B.-** Distribución Binomial: Su parámetro natural es el *logit de “Y”*, es decir, logaritmo de odds de la categoría de aciertos, (la media es una proporción: p/q): $g(\mu_i) = \ln [\mu_i/(1-\mu_i)] = \beta_0 + \sum \beta_i X_i$: Distribución binomial y la función de enlace logit de Y. Se usa este clase de distribución en regresión logística (binaria, nominal, ordinal) y en el análisis de supervivencia (pronósticos en diferente métrica que la “Y”).
- C.-** Distribución Poisson: Su parámetro natural es el logaritmo de la media (μ es un número entero no-negativo). $g(\mu_i) = \ln[\mu_i] = \beta_0 + \sum \beta_i X_i$: Se usa esta distribución para regresión Poisson, modelos log-lineales, (pronósticos en diferente métrica que la “Y”).

Características de los Modelos Lineales

- A.- Modelos Lineales Clásicos (CLM):** ANOVA, Regresión, ANCOVA (llamados modelos lineales generales). Estos modelos son útiles para manejar repuestas cuantitativas y usan función de enlace *Identidad*.
- B.- Modelos Lineales Mixtas (MLM):** ANOVA Model I & Model II, Regrsión multivariable. Estos modelos son útiles para datos de tipo repuestas cuantitativas y utilizan función de enlace *Identidad*. Se diferencian de Clásicos en:
 - B1.-** Incluyen más de un término de error.
 - B2.-** No asumen varianzas iguales ni observaciones independientes.

C.- Modelos Lineales Generalizados (GLM): Incluyen modelo de regresiones de tipo: Poisson, Logística, Binomial, Nominal, Ordinal, Modelos log-lineales y Logit, y modelos del impacto proporcional, etc., y poseen los siguientes rasgos.

C₁- Usan función de enlace logit y logarítmica.

C₂- Respuesta no-cuantitativa.: Se usan en distribuciones dicotómicas (binaria), politómicas (nominal, ordinal), frecuencias.

C₃- Respuestas Cuantitativas que no pueden analizar por los modelos mencionados en C₁ y C₂, por ejemplo modelos de tiempo de supervivencia.

Clasificar modelos en base a las predicciones científicas

Los modelos estadísticos se clasifican en término de las dos predicciones científicas en dos grupos siguientes (Badii et al, 2009).

A.- Modelos para la *predicción I* (analizar diferencia entre datos).

A₁- Modelos en el campo o la investigación *cuantitativa*.

Aquí se trata de 6 modelos que se construyen en base a la relación entre los parámetros (la media o la varianza) y el número de las muestras (varían de 1, 2 y más que 2). Por tanto, estas combinaciones generan los 6 modelos siguientes (Tabla 3.2).

Tabla 3.2. Los modelos en el campo cuantitativo en escala intervalo o razón. Predicción de similitud o diferencia entre grupos de datos (Predicción I).			
Parámetro	Número de muestras		
	2	2	>2
Varianza	X ² (Pearson, 1904)	F (Snedecore, 1934)	(Bartlett, 1937) & Fmax

Promedio	t ₁ de Student (W. Gossett Jr., 1907)	T ₂ de Student (W. Gossett Jr., 1907)	ANOVA (Fisher, 1920)
-----------------	--	--	-------------------------

A2.- Modelos en el campo o la investigación cualitativa.

En función de la escala se generan 2 tipos de modelos (Tablas 3.3 y 3.4).

A2.1.- La escala nominal

Los modelos para esta escala que responden a la predicción I (manejar diferencia entre datos) se presentan en la tabla siguiente (Siegel et al., 1995).

Tabla 3.3. Los modelos estadísticos, en la investigación cualitativa, en escala nominal (a: T.C. = tabla de contingencia).				
Número de muestras				
1	2		>2	
	Independientes	Dependientes	Independientes	dependientes
Ji ² & Binomial Pearson (1904)	T.C. Pearson (1904)	McNemar (1947)	T.C. ^a Pearson (1904)	Q de Cochran (1947)

A2.2.- La escala ordinal

Los modelos para la escala ordinal que responden a la predicción I (tratar de diferencia entre datos) se presentan en la tabla siguiente (Badii et al, 2009).

Tabla 3.4. Los modelos estadísticos, en la investigación cualitativa, en escala ordinal				
Número de muestras				
1	2		>2	
	Independientes	dependientes	Independientes	dependientes
Kolmogorov- Smirnov	Mann-Whitney (1947)	Wilcoxon (1945)	Kruskal-Wallis (1952)	Friedman (1940)

B.- Modelos para la predicción II (analizar relación entre datos).

Los modelos para la escala de intervalo o razón que responden a la predicción II (tratar de la interrelación entre datos) se presentan en la Tabla 3.5 (Badii et al, 2009).

Tabla 3.5. Los modelos estadísticos, en la investigación cuantitativa. Predicción para asociación entre variables (Predicción II).	
Número de las variables	
2	>2
<p>Modelo Lineal: Regresión Lineal Simple</p>	<p>Modelos Lineales: Regresión Lineal Múltiple, Modelos Estructurados Lineales (LISREL, AMOS, EQ, PLS Graph, Smart Graph)</p>
<p>Modelos No lineales: 1.- Potencial 2.- Exponencial 3.- Logística 4.- Asintótica sin umbral 5.- Asintótica con umbral 6.- Logarítmica 7.- Semi-logarítmica 8.- Polinomiales de diferentes ordenes 9.- Curvas periódicas, armónicas, o Serie Furrier</p>	<p>Modelos No-Lineal: 1.- Correlación Canónica 2.- Regresión Logística binaria, continua, categórica 3.- Modelos fractales para análisis espacial 4.- Serie de tiempo para análisis temporal</p>

Calidad de los datos

La calidad de los datos en sí es un prerequisite de suma relevancia para cualquier tipo de análisis estadística. En otras palabras, no importa qué tipo de los modelos uno selecciona para tratar los datos, es decir, de tipo modelos lineales, no-lineales,

paramétricos, no-paramétricos, modelos en escala nominal, ordinal, de intervalo, de tipo discreto. Continuo, univariado, multivariado, etc., si los datos sometidos a cualquier de estos modelos, no están de todo correcto entonces los resultados serán erróneas. Aquí se mencionan algunas consideraciones importantes sobre la calidad de datos que se usarán en los análisis estadísticos (Badii et al, 2009; Wheelan, 2014; Stigler, 2017; Kahneman, 2011; Kraus, 2012; Tversky y Kahneman, 1974; Triola, 2013).

¿Qué representa una población?

Una población es un conjunto *total* de los elementos similares que se requiere estudiarlos y están delimitados por la escala del espacio y el tiempo. Esto significa que cuando se interesa estudiar nociones de la población, en realidad se trata de los datos grandes. Es obvio que por razones de limitaciones de los recursos, se estudia la población por medio del muestreo, y en este contexto en mayoría de los casos se trata de muestreo aleatorio (MA) (Badii y Castillo, 2007, 2009).

Cabe señalar que el muestreo aleatorio abre la puerta a unas técnicas estadísticas muy poderosas, sin embargo, tomar una muestra aleatoria es más difícil de lo que aparece. Hay que ser consciente de que muchos errores estadísticos resultan de aplicar *métodos estadísticos buenos a muestras malas*. Además un punto relevante en el muestreo es en relación al tamaño de la muestra, es decir el número de los elementos contenidos en la muestra. Mayor tamaño de la muestra siempre es mejor, ya que se provoca menor grado de margen de error (L), lo cual conduce a que la muestra sea más representativa de la población. Cuando se trata de la calidad de los datos, un factor muy relevante a tomar en cuenta es los tipos de sesgos asociados con los datos. De hecho hay varios tipos de sesgos en este contexto los cuales se presentan a continuación (Badii et al, 2009; Wheelan, 2014; Stigler, 2017).

Sesgo en la selección

En algunos casos, aún con un tamaño muy grande de la muestra podemos tener sesgo. Un ejemplo histórico es el caso de la elección del presidente de Estados Unidos de Norte América en el año 1936, cuando la disputa era entre Franklin D. Roosevelt y el Sr. London. Se realizaron una muestra del tamaño de 10 millones de personas que tenían

teléfono (y por tanto eran ricos para este tiempo) y eran suscritos a Library Digest. El resultado de la encuesta telefónica de estos 10 millones de personas indicaba que el ganador iba a ser el Sr. London, sin embargo, Sr. Roosevelt resultó el ganador de la contienda. Con una muestra tan grande uno lógicamente esperaba que el resultado de la encuesta (más de 8% de la población) reflejara la realidad. Sin embargo, no fue así y la razón de este fracaso de la predicción no se debe al tamaño de la muestra, sino a la situación de que solo en este tiempo muy poca gente poseía teléfono en su casa y por tanto la opinión de la mayoría abrumadora de las personas que podía votar no se tomó en cuenta y este sesgo en no seleccionar adecuadamente la parte correcta de la población fue la causa de mal representación de los datos. Pero con Muestras buenas a mayor tamaño de la muestra, menor será el margen del error “L” [& vice versa con muestras malas] (ejemplo: suscriptores republicanos RICAS de Library Digest, quien prefieren al Sr. London vs Roosevelt en 1936 con un tamaño de la muestra de 10 millones de gentes con teléfono y carros en 1936].

Sesgo en la publicación

Una tendencia natural de los investigadores en diferentes ramas de las ciencias es poner más énfasis en los resultados positivos que los negativos. Por tanto hay una probabilidad muy alta de publicar los resultados positivos que los negativos y este genera un sesgo en contra de los resultados negativos para su publicación (Stigler, 2017).

Sesgo en la memoria

Se realizaron una encuesta sobre el tipo de la dieta (en años anteriores) de dos grupos de las mujeres de la misma edad, un grupo con cáncer y otro sin cáncer. Curiosamente, las mujeres con el cáncer recordaban que en los años anteriores ellas consumían dietas con alto nivel de grasa. Sin embargo, cuando los datos históricos de la dieta de ambos grupos de las mujeres fueron analizados, se notaron que en realidad, no existía diferencia entre las dietas de los dos grupos. Este tipo de sesgo se le denomina “sesgo de memoria”. Es decir, las mujeres del grupo con cáncer tienden a sesgarse en su memoria con mayor nivel de la grasa en su dieta cuando este era falso (Stigler, 2017).

Sesgo en la sobrevivencia

Durante el estudio en el colegio, el promedio de las calificaciones de los alumnos se incrementa en su valor, normalmente, desde el primer año y paulatinamente hacia el último año del colegio. Este incremento del valor del promedio nos da la impresión de que el nivel de la inteligencia de los alumnos se incrementa con el tiempo, como un reflejo del aumento del valor de promedio de rendimientos de ellos con el tiempo. Sin embargo, la realidad radica en que los alumnos menos inteligentes de forma gradual abandonan el colegio a medida que los años del estudio se avanzan y por tanto, es el abandono de los menos aptos para los estudios que ocasiona el incremento del (Stigler, 2017).

Sesgo en los hábitos humanos

Gente que usa vitaminas son más sanas, precisamente, porque son de Tipo de gentes que toman vitaminas más regularmente.

Se realizó un experimento en la el Departamento de la Salud del estado de Massachusetts en estados Unidos de Norteamérica, en donde se declaraban que las nuevas parejas utilizan pijamas de color púrpura a sus niños. Curiosamente, 20 años después de la publicación de dicha investigación, se notó que el 98% de las personas aceptadas en la Universidad de Harvard usaban pijamas de color púrpura de niño, y solamente, el 3% de las personas en las prisiones en el mismo Estado usaban pijamas de color púrpura en sus niñez.

Es obvio que el color de pijama no tiene ninguna relación con el nivel de la inteligencia de las personas, sin embargo, los padres de dichas personas en este estudio eran de clase de gentes que eran muy disciplinadas, y esto hace que sean diferentes (de forma positiva por su disciplina) de otros padres. (Stigler, 2017). En resumen, de todos los sesgos arriba mencionados, se concluye que *hay que Conseguir datos buenos lo cual no es fácil.*

Una vez que se consideran los sesgos arriba mencionados, hay que buscar que los datos proveen una fuente de comparación. Además de los puntos del encontrar los sesgos e evitarlos y buscar fuentes de comparación, hay que producir y tener disponible una base de datos para el futuro.

Sesgo en la medición

Suponemos que deseamos medir el largo de distancia entre dos países o el largo alrededor de un país a lo largo de costas del mismo (como Inglaterra). Recordar que contrario al caso de un cuarto normal de forma rectangular o cuadrado, con líneas rectas (Geometría Euclidiana, solo con líneas rectas), la situación de la las fronteras de un país como Inglaterra no es lineal, sino curvilínea, por tanto, la medición en este caso depende del instrumento de la medición, es decir, a medida que usamos unidades de mediciones más finas (en milímetro versus en kilómetros), entonces, el resultado es más preciso. Por ejemplo, si la unidad de la medición es de cada 200 km, el largo alrededor de Inglaterra es igual a 2,400 km, y para las unidades de 100 km y 50 km los largos son 2,800 km y 3,400 km, respectivamente. Estas diferencias se deben a las curvaturas en la frontera del país que al usar una medida más grande (en este caso 200 km) los largos de las curvaturas no se toman en cuenta, caso contrario al usar medidas más pequeñas como de 50 km en este ejemplo. El mensaje es lo siguiente: en término general, no tiene sentido de hablar de los valores de un largo medido, sin comentar la escala de la resolución utilizado para las mediciones. Por tanto hablar de unos largos como por ejemplo, 28, 476 o 1,231 sin mencionar las unidades de medición es como mezclar centímetro con kilometro o angstrom.

Relevancia de datos y la estadística

Debemos enfatizar el punto crucial que la estadística no comprueba nada con certeza. De hecho el poder de inferencia estadística radica en los puntos siguientes (Badii et al, 2009; Wheelan, 2014). En este contexto se debe enfatizar los temas siguientes.

- A.-** Observar un patrón o resultado.
- B.-** Usar la probabilidad para determinar la explicación más probable. [debemos recordar que, obviamente, la explicación más probable, no es siempre la correcta].
- C.-** Con respecto a la inferencia estadística: Se trata de tomar decisiones y sacar conclusiones. En este contexto hay que enfatizar los siguientes tres puntos.
 - C1.-** Procesos a través de los cuales comunicamos con datos para generar

conclusiones significativas.

C2.- Es crucial notar que la unión entre los dos conceptos, es decir, los datos y la probabilidad (con el apoyo de “Teorema de Límite Central”) nos conducen a la inferencia correcta.

C3.- Nos permite tener una representación correcta del universo a partir de la muestra y con el apoyo de la Teorema de Límite Central (TLC).

D.- En relación a las pruebas de hipótesis: Las hipótesis nula (H_0) y la alterna (H_a) son complemento lógicos, esto significa que si rechazamos una, entonces, debemos aceptar la otra. También hay recordar algo que parece contra-intuitivo, es decir, los investigadores frecuentemente, crean H_0 con la esperanza de rechazarla.

Métricas y claves para la calidad de los datos (Badii et al, 2009).

Existen diversas métricas para colección de los datos, Sin embargo, el reto real es buscar “*Datos de Calidad*”. Una clave o orientación en este contexto es la “**COETE**” La COETE es una base conceptual que apoya a distinguir la naturaleza correcta de datos a seleccionar. Las 5 siglas de la palabra COETE se tratan de lo siguiente:

Componente: Individuo, pandilla, población, comunidad, fuerza militar, empresa.

Operación: Cooperación, Confrontación (intencional, no-intencional), Comensal, Amensal.

Escala: Tiempo, espacio.

Tipo de datos: Anécdota, investigación, documental, observación del campo, experimental.

Esencia: Lineal, curvilínea (de diferentes tipos), circular con o sin umbral.

Teoría general de Sistemas (TGS): Acercamiento

Hay que recordar que la sociedad es como un sistema. En este contexto, el comportamiento de la sociedad se determina vía interacción de sus partes y con su entorno. El Sistema tiene un objetivo y todos los partes operan para alcanzar este objetivo. Según la

Segunda Ley de Termodinámica: La tendencia natural es ir hacia el desorden (la Entropía), sin embargo, la tendencia del sistema es el encontrar el equilibrio interno (obviamente, al costo del desorden externo).

Cuando se trata de la Teoría General de Sistemas, hay que recalcar los puntos siguientes.

A.- TGS requiere de una *comunicación* eficaz y eficiente.

B.- Relación *causal* en el sistema: un cambio en una parte genera cambios en *todo* el sistema.

C.- Tendencia innata del sistema para mantener su *equilibrio interno*.

D.- La *longevidad* del sistema depende en su capacidad de *adaptación* a los cambios.

Conclusión

Intuición es una guía inadecuada para la búsqueda de la verdad. Se requiere de observar, planear, establecer hipótesis, ejecutar el plan, evaluar los resultados para llegar con realidades espacio-temporales. En el contexto de la ejecución, la forma más certera es el realizar experimentación científica y aun así, todos los experimentos están asociados con error (error experimental), debido a la noción de que la experimentación se ubica dentro de la lógica inferencial. Por tanto, el sesgo, el error y el concepto de aleatoriedad son partes esenciales en las investigaciones científicas. Solamente con una comprensión correcta de la estadística, leyes de probabilidad y el empleo de la racionalidad, se puede mitigar los errores asociados.

Referencias

- Badii, M.H. y J. Castillo (eds). 2007. Técnicas Cuantitativas en la Investigación. UANL, Monterrey.
- Badii, M.H. y J. Castillo. 2009. Muestreo Estadístico: Conceptos y Aplicaciones. UANL, Monterrey.
- Badii, M.H., J. Castillo & A. Wong. 2006. Diseños de distribución libre. Innovaciones de Negocios, 3(1): 141-174.

- Badii, M.H., J. Castillo, J. Rositas y G. Alarcón. 2007. Uso de un método de pronóstico en investigación. Pp. 137-155. In: M.H. Badii & J. Castillo (eds.). Técnicas Cuantitativas en la Investigación. UANL, Monterrey.
- Badii, M.H., J. Castillo, J. Landeros y K. Cortez. 2009. Papel de la estadística en la investigación científica. Pp. 1-43. In: M.H. Badii & J. Castillo (eds). Desarrollo Sustentable: Métodos, Aplicaciones y Perspectivas. UANL. Monterrey.
- Kahneman, D. 2011. Thinking Fast and Slow. Farrar, Straus and Giroux, New York. 499 p.
- Celis, A.J. y V. Labrada. 2014. Bioestadística 3ª. Ed. Manual Moderno, México, D.F. 337 p.
- Kepner, J.L. y D.H. Robinson, 1988. Nonparametric methods for detecting treatment effects in repeated measures designs. J. Amer. Statist. Assoc. 83: 456-461.
- Keppel, G. 1991. Design and Analysis: A Researcher's Handbook 3rd ed. Prentice Hall. Englewood Cliffs New Jersey.
- Kirk, R.E. 1982. Experimental Design: Procedures for the Behavioral Sciences. 2nd ed. Brooks-Cole, Monterey California.
- Kraus, J. 2012. Uncertainty. 101-102. En: J. Brockman (ed). 2012. This Will Make You Smarter, Harper Perennial, 540 p.
- Morris, T.R. 1999. Experimental Design and analysis in Animal Sciences. CABI Publishing, London.
- Outhwaite, A.D. y A. Rutherford, 1955. Covariance analysis as alternative to stratification in the control of gradients. Biometrics, 11: 431-440.
- Pezzullo, J.C. 2013. Biostatistics. John Wiley & Sons, Hoboken, NJ. 390 p.
- Samuels, M.L., G. Casella, and G.P. McCabe. 1991. Interpreting blocks and random factors. J. Amer. Statist. Assoc. 86: 798-808.
- Siegel, S. y N.J. Castellum. 1995. Estadística No Paramétrica Aplicada a las Ciencias de la Conducta. Trillas, México, D.F.
- Sokal R.R. and F.J. Rohlf. 1969. Biometry The Principle and Practice of Biostatistics in Biological Research.. W.H. Freeman and Company, San Francisco. 776 p.
- Stigler, S.M. 2017. Los Siete Pilares de la Sabiduría Estadística. G. S., Cd. de México. 181 p.
- Triola, M.F. 2013. Estadística. 11a. ed. Pearson, México, D.F. 856 p.
- Tverky, A. and D. Kahneman. 1974. Judgement under uncertainty: Heuristics and biases. Science, New Series, 185(4157): 1124-1131.
- Wheelan, C. 2014. Naked Statistics. Stripping the Dread from the Data. W.W. Morton & Company, New York. 282 p.